Multi-omics facilitated variable selection in Cox-regression model for cancer prognosis prediction

Cong Liu[1,4], Xujun Wang[2,3], Georgi Z. Genchev[2,4], Hui Lu[1-4,*]

1. Department of Bioengineering, University of Illinois at Chicago, Chicago, U.S.A
2. SJTU-Yale Joint Center for Biostatistics, Shanghai Jiaotong University, Shanghai, China
3. Department of Bioinformatics and Biostatistics, Shanghai Jiaotong University, Shanghai, China
4. Center for Biomedical Informatics, Shanghai Children's Hospital, Shanghai, China

*Corresponding author

Email Address:

CL: cliu55@uic.edu

XW: jameswong.tju@gmail.com

GG: george.genchev@gmail.com

HL: huilu@sjtu.edu.cn

Abstract

Motivation: New developments in high-throughput genomic technologies have enabled the measurement of diverse types of *omics* biomarkers in a cost-efficient and clinically-feasible manner. Developing computational methods and tools for analysis and translation of such genomic data into clinically-relevant information is an ongoing and active area of investigation. For example, several studies have utilized an unsupervised learning framework to cluster patients by integrating *omics* data. Despite such recent advances, predicting cancer prognosis using integrated *omics* biomarkers remains a challenge. There is also a shortage of computational tools for predicting cancer prognosis by using supervised learning methods. The current standard approach is to fit a Cox regression model by concatenating the different types of *omics* data in a linear manner, while penalty could be added for feature selection. A more powerful approach, however, would be to incorporate data by considering relationships among *omics* datatypes.

Methods: Here we developed two methods: a SKI-Cox method and a wLASSO-Cox method to incorporate the association among different types of *omics* data. Both methods fit the Cox proportional hazards model and predict a risk score based on mRNA expression profiles. SKI-Cox borrows the information generated by these additional types of *omics* data to guide variable selection, while wLASSO-Cox incorporates this information as a penalty factor during model fitting.

Results: We show that SKI-Cox and wLASSO-Cox models select more true variables than a LASSO-Cox model in simulation studies. We assess the performance of SKI-Cox and wLASSO-Cox using TCGA glioblastoma multiforme and lung adenocarcinoma data. In each case, mRNA expression, methylation, and copy number variation data are integrated to predict the overall survival time of cancer patients. Our methods achieve better performance in predicting patients' survival in glioblastoma and lung adenocarcinoma.

Key words: *multi-omics*; variable selection; cancer prognosis prediction; Cox regression

Introduction

Consideration of histological assays and population-based risk factors (such as family history, behavior, age, etc.), combined with environmental risk factors – such as assessment of exposure to environmental carcinogens - are commonly used in clinical settings in an effort to determine cancer prognosis and patient outcomes [1]. Advances in molecular biology and high-throughput technology in the last two decades have precipitated the availability of new tools in the diagnostic and prognostic armament by enabling the simultaneous measurement of vast number of biomarkers in a single experiment. In a *multi-omics* landscape, for a single patient, it is well-possible for multiple types of genome-scale data such as mRNA expression, copy number variants (CNV), and methylation to be collected.

Such large scale data-focused approaches have been used to predict cancer prognosis using mRNA expression profiles [2,3]. Several cancer survival-associated expression biomarkers have been identified. A panel of three genes (MAMDC2, TSHZ2, and CLDN11) were identified as significantly correlated with survival of breast cancer patients [4]. More recently, research groups have established a Cox regression model incorporating the expression of IL6, IL1A, and CSF to predict survival of diffuse large B-cell lymphoma patients [5]. Besides mRNA expression, other molecular markers also show power in predicting cancer patients outcomes. For example, the PITX2 DNA methylation has shown prediction capability for prostate cancer survival [6]. Another study suggests that patients with a higher expression of microRNA-155 had significantly worse recurrence-free survival [7] and CNV have been also linked to cancer prognosis in several studies [8,9].

A common analytical task is to link the measurement of these genomic covariates to the patients' survival time or time to cancer relapse, which is usually censored data. A popular strategy is to fit a Cox regression model using these covariates for censored survival data, and then predict the cancer prognosis for a new patient based on this fitted model [10-12]. The high-dimensionality issue i.e. when p (dimension of the data) $\gg$ n (number of observation), introduced by high-throughput data increases the difficulty of downstream analysis [2]. To reduce the high-dimensionality, variable selection is a common procedure in predicting the prognosis of cancer. LASSO and its variants (e.g. Adaptive LASSO, elastic-net, etc.), are a popular strategy to provide variable selection in regression analysis and have been extended to Cox regression model [13-15].

Briefly, the LASSO-Cox estimators maximize the Cox partial likelihood with an L-1 constraint on coefficients. To predict prognosis using LASSO-Cox, the simplest way is to concatenate measurements from various *omics* levels. Unfortunately, concatenation will further increase the p, making the high-dimensionality issue worse. Moreover, concatenation ignores the potential association between different levels of *omics* data. For example, strong correlation between DNA methylation and mRNA expression has been found in various diseases [16]. Therefore, to maximize the utility of information from different *omics* levels, sophisticated strategies for variable selection of *multi-omics* data should be designed.

Herein we propose two novel methods for variable selection in cancer prognosis prediction. Unlike traditional concatenation methods, we only use expression data to train and predict in a Cox regression framework, while other *omics* data is used for variable selection. The rationale for our method is firstly, not to dilute the already low signal-to-noise ratio in one data type, and secondly, to take the link among different types of data into account to assist variable selection. Utilizing other *omics* data, that is mapped at the gene level, we could draw additional marginal partial correlations that could be further summarized and integrated into a single vector representing the correlation between gene and survival. The first method, SKI (screen with knowledge integration)-Cox, based on our previous work [17], first screens genes based on the knowledge derived from other *omics* data, and then fits a LASSO-Cox model by mRNA expression profiles of selected genes. The second one, wLASSO (weighted LASSO)-Cox borrows the idea from Adaptive LASSO-Cox model [18]. In this approach, the penalty factor for each coefficient is adjusted by coefficients obtained by fitting survival time with other (such as methylation, CNV, etc.) *omics* data.

This paper is organized as follows: First we briefly review the LASSO-Cox regression and present our developed methods. We then evaluate the performance of each proposed method by simulation studies and applications to two TCGA (The Cancer Genome Atlas) datasets, GBM (glioblastoma) [19] and LUAD (lung adenocarcinoma) [20]. Finally, we discuss our methods and their utility in clinical applications.

Methods

SKI-Cox and LASSO-Cox

Suppose we have a sample size of n patients: $(y_1, \delta_1, X_1), (y_2, \delta_2, X_2), \dots, (y_n, \delta_n, X_n)$, where $y_i = \min(t_i, u_i)$ is the observed time (i.e. time to the death t_i, or time to the last follow up u_i), δ_i is the censoring indicator (i.e. $\delta_i = 1$ if $t_i \leq u_i$), and $X_i \in R^p$ is the *omics* measurement (e.g. mRNA expression profile, methylation profile, etc.). The Cox proportional hazard model assumes the hazards (or instantaneous death rate) at time t :

$$\lambda(t; X_i) = \lambda_0(t) \exp\,(X_i\beta)$$

where $\lambda_0(t)$ is the baseline hazard, and $\beta = \{\beta_1, \beta_2, \ldots, \beta_p\}$ is the vector of regression coefficients.

The partial likelihood defined by Cox is:

$$L^{par}(\beta; X) = \prod_{i=1}^{n} \left[\frac{\exp(X_i\beta)}{\sum_{j \in R(y_i)} \exp\left(X_j\beta\right)}\right]^{\delta_i}$$

where $R(t) = \{i : y_i \geq t\}$ denotes the set of individuals “at risk” for death at time t.

In Cox regression, the estimators are obtained by maximizing the partial log likelihood. When the dimension of β increases, LASSO estimators are often used to introduce sparsity by maximizing the L_1-penalized partial log likelihood:

$$\beta^{LASSO} = \arg\max \frac{1}{n} l^{par}(\beta; X) - \lambda \sum_{j=1}^{p} |\beta_j|$$

In the *multi-omics* case, the predictor X_i becomes $\{X_i^{(1)}, X_i^{(2)}, \ldots, X_i^{(K)}\}$, where each $X_i^{(k)}$ represents a single data type with dimension p_k for each patient i. The concatenation procedure will combine all the $X_i^{(k)}$ into a concatenated vector X_i with dimension $\sum_{k=1}^{K} p_k$.

We observed that such concatenation method will worsen the high-dimensional issue, and furthermore, due to the existence of high correlation among different levels of *omics* data, the LASSO procedure could be unstable [21]. To overcome that, some research groups have argued that only variables with the most direct effect on cancer clinical outcomes such as mRNA expression should be used, while other measurements (such as methylation) that affect outcomes through regulating mRNA expression could be ignored [22]. Our approach agrees partially with this argument by fitting LASSO-Cox regression model using only mRNA expression data. However, we believe that data from other *omics* types could potentially improve the variable selection procedure. Therefore, we implement two methods to facilitate variable selection using *multi-omics* information.

In essence, our approach is as follows: let's suppose that we have mRNA expression data, methylation data and CNV data profiles available. All data types are first standardized and then the target values are calculated at the gene level. For example, in the case of Illumina 450K-array based methylation data, a gene methylation value could be calculated by taking the mean signal values of probes for each gene [23]. Finally, for each patient i and each gene j, $X_{ij} = \{X_{ij}^{(expr)}, X_{ij}^{(cnv)}, X_{ij}^{(methyl)}\}$, where $X_{ij}^{(expr)}, X_{ij}^{(cnv)}, X_{ij}^{(methyl)}$ are the values for mRNA expression, copy number variance and methylation, respectively. The idea of SKI-Cox is based on variable screening in Cox's model [24]. First, a single-covariate Cox regression model for each gene j at each *omics* type k, is fitted and the marginal utility U_j^k:

$$U_j^{(t)} = \max_{\beta_j} \frac{1}{n} l^{par}(\beta_j; X^{(k)})$$

, defined as the maximum of the partial likelihood of gene j is obtained. Once we have obtained all marginal utilities $U_j^{(k)}$ for $j = 1,2,\dots,p$, we could rank all covariates at k^{th} *omics* type by corresponding marginal utility in descending order. Following that, we combine ranks at different *omics* types to generate a novel rank as:

$$R_j = rank\left(\frac{1}{3}(R_j^{(expr)} + R_j^{(cnv)} + R_j^{(methyl)})\right)$$

Though average ranks were used in our approach, weighted average or other function (e.g. min, max, median, etc.) could be tried or learned from data if the observation set is big enough. Subsequently we selected the top *d* ranked (e.g. 2000) covariates and denote $\mathcal{M}$ as the index set of these *d* covariates. We acknowledge that the number of variables selected in this step is subjective and heuristic, better strategy [25] could be used to determine this number. However, it is out of the scope of this study. We further fitted a LASSO-Cox regression model using $X_{\mathcal{M}}^{(expr)}$.

$$\beta^{SKI} = \arg\max \frac{1}{n} l^{par}\left(\beta; X_{\mathcal{M}}^{(expr)}\right) - \lambda \sum_{j=1}^{p} |\beta_j|$$

and wLASSO-Cox implementing an extended LASSO-Cox regression model.

$$\beta^{W} = \arg\max \frac{1}{n} l^{par}\left(\beta; X^{(expr)}\right) - \lambda \sum_{j=1}^{p} \left|\beta_j\right| \tau_j$$

where $\tau_j = \frac{3}{\left|\tilde{\beta}_j^{(expr)}\right| + \left|\tilde{\beta}_j^{(cnv)}\right| + \left|\tilde{\beta}_j^{(methyl)}\right|}$, and $\tilde{\beta}_j^{(k)}$ is the maximizer of the log partial likelihood of a single covariate $l^{par}\left(\beta_j; X^{(k)}\right)$. The logic behind it is similar to Adaptive LASSO. If the coefficients carry more information across different *omics* types; then less penalty will be applied. However, in adaptive LASSO case, the penalty factors are obtained by fitting a full log partial likelihood, while in our case, the information contributed for a same variable across different *omics* types might not be consistent (i.e. the coefficient values could be different). Therefore, we select the maximizers of the marginal log partial likelihood as our penalty factors.

Simulation methodology

The performance of our proposed methods were first tested in a simulation study. We focused on whether using the information brought by *multi-omics* data could help select important covariates in a *single-omics* (e.g. mRNA expression) model. In our simulation, we generated 500 patients, and 20,000 covariates (i.e. mRNA expression). We assumed the first 20 covariates were related to the cancer prognosis (i.e. survival time) through a Cox hazards proportional model. 10 coefficients were generated from the following uniform distribution: $Unif(-1,-0.1)$, and another 10 were generated from: $Unif(0.1,1)$.

For the values of covariates, we first generated a $500\times 20{,}000$ dataset $X^{(1)}$ from $Unif(-1.5,1.5)$. In order to mimic the possible interaction among genes, we used Gram–Schmidt ortho-normalization [26] to construct its normalized orthogonal basis $A = \{a_1, a_2, \dots, a_{20}\}$ and $B = \{b_1, b_2, \dots, b_{480}\}$, where A is the linear space expanded by ${X_{1:20}}^{(1)}$. We then generated the expression levels for the rest of the genes from the linear space $C = B + AE$, where $E \in R^{20\times 480}$ could be any matrix. In our simulation, we set the values in E all equal to a single number e. We note when $e = 0$, then the expression of the rest of the genes are independent of the first 20 genes. The correlation could be adjusted by controlling the size of e. Three e's (0.1, 0.2, 0.3) were generated to represent different strength of correlations.

The survival times were then generated from a Weibull distribution with shape $v = 5$ and scale

$\lambda = 5$. The time was generated according to [27,28]:

$$t_i = \left(-\frac{\log(u)}{\lambda \exp(X_i \beta)}\right)^{\frac{1}{v}}$$

where $u \sim Unif(0,1)$.

The censoring times u_i were generated from $Unif(2,10)$, and we then generated the observation time $y_i^{(1)} = \min(t_i, u_i)$. Based on above setting, we could expect a censoring rate is about 40%.

We also generated another three datasets - $\{X^{(2)}, y^{(2)}, \delta^{(2)}\}$, $\{X^{(3)}, y^{(3)}, \delta^{(3)}\}$, and $\{X^{(4)}, y^{(4)}, \delta^{(4)}\}$ to represent other types of *omics* data using the same procedure described above. Two settings were considered. In the first setting, the first 20 coefficients were set as non-zero in all data types, though the values of coefficient for the same covariate could be different across different types of *omics* data. In the second one, in each data type except for $\{X^{(1)}, y^{(1)}, \delta^{(1)}\}$, 12 of the first 20 coefficients were randomly sampled and set as nonzero, and the rest 8 were set as zero.

Data processing and performance evaluation in TCGA data application

To further demonstrate the utility of our proposed methods, we applied them to predict TCGA patient survival data. Processed clinical data, mRNA-sequencing based expression data, Illumina 450K based methylation data, and SNP-array based copy number variance data were downloaded through the FirebrowseR provided by TCGA consortium [29]. For a given gene, the expression value was represented as $log(RSEM)$, where $RSEM$ estimate the relative transcripts abundance by effectively using ambiguously-mapped reads [30]. Methylation values were summarized as the mean value of all the probes annotated within this gene. CNV data were first processed by GISTIC2.0 [31], and then the value of a specific CNV segment, representing amplification and deletion status, was assigned to each gene located within its genomic region. More sophisticated strategy could be applied or even learned from the data to infer these mapping relationship [32]. However, more parameters are introduced if we employ this strategy, and this could further complicate the problem. Thus this heuristic-based and easy-to-implement mapping strategy was used here.

We used mRNA expression data to predict patient overall survival with variable selection facilitated by methylation and CNV data. We used *C-index* with 10-fold cross validation to evaluate the performance. Briefly, *C-index* measures the fraction of patient pairs, where the observation with the higher survival time have the higher survival score predicted by the models [33]. The *concordance.index* function in the R package *survival* was used for implementation. Survival scores were calculated as a linear combination of coefficients and covariates:

$$score_i = -X_i\hat{\beta}$$

Results

Our proposed methods successfully recovered more true variables in the simulation study by incorporating the information from other *omics* data, and achieved better performance in predicting patients' survival in glioblastoma and lung adenocarcinoma.

Variable selection in simulation study

To see whether our proposed method could use other *omics* data to improve the selection of true variables in a *single-omics* dataset, we compared our methods with a regular LASSO-Cox based method which only uses $\{X^{(1)}, y^{(1)}, \delta^{(1)}\}$ (e.g. mRNA expression). To make the results comparable, we modified the tuning parameters until 200 variables were selected in each method. **Table 1** shows the number of true variables selected using the three methods under different settings. When the correlation between true covariates and others are low, all three methods perform very well. Our proposed methods perform slightly better than a LASSO-Cox regression. With the increase of the correlation, the capability of LASSO-Cox to select true covariates drop dramatically, while the false positive rate of SKI-Cox and wLASSO-Cox could still remain at a relatively low level. When the maximum correlation increases to about 0.4, the task becomes extremely difficult due to strong collinearity among the covariates. LASSO-Cox could not find even one true variable, while SKI-Cox and wLASSO-Cox were able to identify at least one true variable.

We also simulated the situation when the different types of *omics* data do not predict the prognosis consistently. When the shared proportion of informative variables between *omics* data drops to 60%, we observed the performance of SKI-Cox and wLASSO-Cox decline when the task is easy (i.e. low correlation). However, we observed that the number of true positive rate of SKI-Cox and wLASSO-Cox could still exceed LASSO-Cox when the task becomes more and more complicated (i.e. high correlation). Overall, wLASSO-Cox tends to perform slightly better than SKI-Cox.

Predictions in glioblastomas (GBM) and lung adenocarcinoma (LUAD)

We then applied our models in two cancer dataset in TCGA, lung adenocarcinoma

(LUAD) and glioblastomas (GBM). The overall survival time is longer for lung adenocarcinoma patients than that for glioblastomas patients, and thus results in a higher censoring rate in LUAD dataset. We also observe high correlations among genes at each *omics* type. From the learnings in our simulation, it is very likely that we could not find a true signal when a high correlation exists. Therefore, we focused our comparison on survival prediction. The clinical information and overall *omics* information were summarized in **Table 2**.

We compared our methods with four other methods. LASSO-Cox^{expr} uses only expression data to predict survival time, LASSO-Cox^{cnv} uses only CNV data, LASSO-Cox^{methyl} uses only methylation data, and LASSO-Cox^{concat} first concatenates the *omics* data and then fits a LASSO-Cox regression model to predict survival using the concatenated matrix. In general, the performance is better in LUAD than the performance in GBM, which is likely due to that fact that more patients have *omics* data available in LUAD. In the *single-omics* case, mRNA expression achieves a better performance (average *C-index* 0.53 and 0.58) than methylation (average *C-index* 0.51 and 0.51) and CNV (average *C-index* 0.51 and 0.56).

In both cancers, using CNV and methylation *omics* data by simple concatenation do not bring additional predictive power. Instead, the prediction performance declines (average *C-index* 0.51 and 0.57) likely due to the extra noise introduced by the additional *omics* data. However, our proposed methods do improve the prediction in both cancers by introducing the other (methylation and CNV) *omics* data in variable selection procedures. In GBM, SKI-Cox (average *C-index* 0.62) works better than wLASSO-Cox (average *C-index* 0.59), while wLASSO-Cox performs better in LUAD (average *C-index* 0.60 vs. 0.63).

To identify the variables selected by different methods, we used bootstrap (100 times) to show the most frequently selected genes (**Table 4** and **Table 5**). In both GBM and LUAD, the most frequently selected variables using concatenated LASSO-Cox^{concat} are from mRNA expression data, which further confirms the assumption that mRNA expression, which has the most direct impact on phenotypes, could have the most predictive power. The variables selected in LASSO-Cox^{concat} and LASSO-Cox^{expr} are very similar despite the frequency is higher in LASSO-Cox^{expr}, which is likely due to the more variables in LASSO-Cox^{concat}. Comparing to CNV and methylation based model, the frequency of specific most often selected variables is higher in the model consisting of mRNA expression, indicating a higher model stability when we use expression based data. The stability increases even more in our proposed two models. For example, the expression of STXBP4 and MBLAC2 have been frequently selected in GBM as the predictive variables, and their selection frequency is much higher in our proposed models. Moreover, with the addition of information from methylation and CNV, some genes not selected before will show their predictive powers (ARPC1A and INHA). mRNA expression of INHA, a tumor suppressor, was altered in adrenocortical carcinoma patients (ACC) [34]. The alteration of ARPC1A, which is another tumor suppressor, has also been observed in multiple cancers progress including GBM [35,36]. This fact underscores the capability of our approach to discover potentially clinically-useful biomarkers not captured by other models.

Discussion

Application of *multi-omics* data based approaches towards the goal of informing patient-focused decision making has gained popularity in recent years. Several methods have utilized *multi-omics* data to perform patient clustering. iCluster developed a joint Gaussian latent variable model for integrated *multi-omics* clustering [37]. Subtypes showing poor survival were discovered by applying iCluster algorithm in breast and lung cancers using mRNA expression and copy number data. SNF used network fusion techniques to build patient similarity network by integrating mRNA expression, DNA methylation and microRNA (miRNA) expression data. Survival risk could be predicted using a Cox regression model with penalty applied to control the patient similarity [38]. PARADIGM inferred the pathway activity using *multi-omics* data and clustered patients based on these activities [39]. Most of these approaches require all types of *omics* data available for both training and prediction.

Unlike such methods, our methods do not incorporate all data types into one model. Instead, we only used mRNA expression as a basic data type to train the model, while other types of *omics* data were only used to facilitate variable selection. In the SKI-Cox approach, variables (i.e. genes) are first screened from different genomic points of view based on *omics* data types, and then ranked based on their average marginal utility in survival prediction. The final model is trained using the mRNA expression data of these screened genes. The other model, wLASSO-Cox puts a penalty factor to take the information derived from other (CNV and methylation) *omics* data into account. The more predictive power a gene shows in other *omics* data, the less penalty it has in an mRNA expression-based regression model. The idea of SKI was first developed in our previous study [17], in which informative variables were first screened based on prior biological knowledge. In this current application, *multi-omics* data could be regarded as a layer of knowledge. Similarly, wLASSO extended the idea of Bergersen and his colleagues' work [40], in which prior knowledge was integrated into a LASSO regression model as a penalty factor. However, both of the previous works only considered a simple linear regression model. In our case, we extended the work to a Cox regression model.

An obvious advantage of our approach is that different *omics* data could be measured in different patients. For example, we could have methylation data measured for one group of patients to derive the predictive power of each gene from methylation perspective, and then apply this to another group of patients to train an mRNA expression-based Cox regression model. Since mRNA expression is the most commonly measured genome-scale marker in clinical applications, such a model setting allows us to collect more training samples, which could be essential when handling the multi-dimensionality ($p >> n$) issue. On the other hand, a prediction based on our model only requires the sample to be measured in a single *omics* level (e.g. mRNA expression). Considering the still-high price to measure genomic-scale data and relative small amounts of biopsy materials available for measurement, our methods could maximize its utility in clinical applications. The reason we selected mRNA expression to train the final model is inspired by the observation that mRNA expression has higher predictive power than the other genomic measurements, which is an expected result since mRNA expression has the most direct effect on cancer clinical outcomes [22]. Furthermore, as the most mature genome-scale technology, mRNA-expression is the most popularly applied clinical tool to measure genomic-scale data, which could make our methods more widely adopted and useful in the clinical setting.

Besides the prediction tasks, our methods enjoy the sparsity property as a result of LASSO-based regression. The final variables selected in our model (e.g. $\beta \neq 0$) could be downstream-validated and designed as a gene panel for future clinical usage. Since we incorporated other *omics* data in variable selection, it is more likely the final variables are those driver genes, due to the fact that the upstream regulators of these genes also show predictive power in survival prediction. Here our assumption is that the "signal" is sparse. It is possible that many genomic features could contribute to the cancer prognosis. Then other feature reduction methods such as

PCA (principle component analysis) [41] and PLS (partial least squares) [42]should be implemented in a Cox-regression framework.

In conclusion, we have developed two methods SKI-Cox and wLASSO-Cox to facilitate variable selection in Cox-regression model using *multi-omics* data. The performance has been validated by both simulation and real case studies. More true variables could be recovered in the simulation study. Better performance is achieved in predicting overall survival time in glioblastoma and lung adenocarcinoma patients using TCGA dataset. Our methods introduce a novel framework for variable selection in Cox-regression model using multi-omics data. Its easy-to-implement property makes it very promising and useful in building a clinically applicable predictive model. The procedure we applied could also help identify driver genes and shed the light in explaining cancer development, prognosis, and relation to patient-specific outcomes.

Declarations

Author contributions

CL designed the concept, derived statistical methods, collected the application data, wrote the programming code, perform the data analysis and drafted the manuscript; WJ: derived the statistical model and collected the application data; GG: designed the concept and drafted the manuscript; HL: designed the concept, provided financial support, and approved the final manuscript.

Competing interest

The authors declare that they have no competing interests.

Funding

Publication of this article was funded by UIC and by the National Natural Science Foundation of China (No.31071167 and No. 31370751)

	t	MAC	β Overlap	# of True variables (among 200)
LASSO-Cox	0.1	0.132	1	18.7
SKI-Cox	0.1	0.132	1	19.9
wLASSO-Cox	0.1	0.132	1	20
LASSO-Cox	0.1	0.132	0.6	18.8
SKI-Cox	0.1	0.132	0.6	14.5
wLASSO-Cox	0.1	0.132	0.6	16.1
LASSO-Cox	0.2	0.256	1	6.3
SKI-Cox	0.2	0.256	1	10.9
wLASSO-Cox	0.2	0.256	1	12.1
LASSO-Cox	0.2	0.256	0.6	6.4
SKI-Cox	0.2	0.256	0.6	6.6
wLASSO-Cox	0.2	0.256	0.6	7.9
LASSO-Cox	0.3	0.411	1	0
SKI-Cox	0.3	0.411	1	1.2
wLASSO-Cox	0.3	0.411	1	3
LASSO-Cox	0.3	0.411	0.6	0
SKI-Cox	0.3	0.411	0.6	0.3
wLASSO-Cox	0.3	0.411	0.6	0.8

Table 1. Simulation results showed number of true non-zero β variables selected using three different methods under different scenarios. MAC: maximal absolute correlations among variables.

		GBM	LUAD
Clinical outcomes	Number of patients	591	509
	Average overall survival (month)	501.0+	909.9+
	Event rate	82.91%	35.95%
Omics summary	# of genes measured	18,218	18,309
	# of patients with mRNA expression	151	390
	# of patients with methylation	280	333
	# of patients with CNV	554	389
	MAC among mRNA expression	0.98	0.97
	MAC among methylation	0.94	0.95
	MAC among CNV	1	1

Table 2. Clinical and *omics* data summary of GBM and LUAD. MAC: maximal absolute correlation; LUAD: lung adenocarcinoma; GBM: glioblastomas.

Method	*C-index* (standard error)	
	GBM	LUAD
LASSO-Cox^{expr}	0.53 (0.02)	0.58 (0.03)
LASSO-Cox^{cnv}	0.51 (0.02)	0.56 (0.01)
LASSO-Cox^{methyl}	0.51 (0.01)	0.51 (0.02)
LASSO-Cox^{concat}	0.51 (0.03)	0.57 (0.03)
SKI-Cox^{expr}	0.62 (0.01)	0.60 (0.01)
wLASSO-Cox^{expr}	0.59 (0.02)	0.63 (0.02)

Table 3. *C-index* obtained by 10-fold cross-validation of different methods. LUAD: lung adenocarcinoma; GBM: glioblastomas.

LASSO-Cox-$_{\text{concat}}$	LASSO-Cox$^{\text{expr}}$	LASSO-Cox$^{\text{methyl}}$	LASSO-Cox$^{\text{cnv}}$	SKI-Cox$^{\text{expr}}$	wLASSO-Cox$^{\text{expr}}$
STXBP4$^{\text{expr}}$ 0.75	STXBP4 0.78	USP49 0.45	ZC3H12C 0.45	STXBP4 1.00	STXBP4 0.99
ARHGAP42$^{\text{expr}}$ 0.38	SH2D6 0.47	FAM3B 0.41	RDX 0.40	MBLAC2 0.89	MBLAC2 0.89
FAM3B$^{\text{methyl}}$ 0.37	HLA-DRB6 0.45	LRRC8E 0.37	AHDC1 0.39	ARPC1A 0.74	LIMA1 0.84
SH2D6$^{\text{expr}}$ 0.33	NSUN5 0.40	CAB39 0.35	FGR 0.33	C11orf35 0.69	TXN 0.82
MBLAC2$^{\text{expr}}$ 0.27	MBLAC2 0.34	A4GALT 0.31	R3HDM2 0.29	USP6NL 0.61	TMEM44 0.68
FAHD2A$^{\text{expr}}$ 0.24	CUL5 0.33	GDNF 0.28	AKAP6 0.29	C19orf73 0.43	ARPC1A 0.65
RPS28$^{\text{expr}}$ 0.22	GPR126 0.33	CYB5R3 0.25	FDX1 0.28	INHA 0.40	INHA 0.64
SLC2A2$^{\text{methyl}}$ 0.21	NFXL1 0.33	AGPAT1 0.22	ACSM3 0.28	CPNE2 0.36	B4GALT5 0.51
CUL5$^{\text{expr}}$ 0.19	ARHGAP42 0.32	PIK3IP1 0.22	LINC00290 0.26	C21orf2 0.35	HEY1 0.37
GPR126$^{\text{expr}}$ 0.19	FAM35A 0.30	FUT9 0.21	FAM138F 0.26	MAML2 0.32	DCAF17 0.31

Table 4. 10 most frequently selected variables of GBM in different models using bootstrap methods. *expr*: mRNA expression features; *methyl*: methylation features.

LASSO-Cox-^{concat}	LASSO-Cox^{expr}	LASSO-Cox^{methyl}	LASSO-Cox^{cnv}	SKI-Cox^{expr}	wLASSO-Cox^{expr}
PLEKHB1^{expr} 0.80	PLEKHB1 0.82	ZFAND2A 0.80	SEPT14 0.43	MYLIP 0.88	MYLIP 1
MYLIP^{expr} 0.77	MYLIP 0.80	PDZD8 0.72	MRPS17 0.39	FUT4 0.86	ZNF330 0.92
RGS20^{expr} 0.59	FAM117A 0.59	XPA 0.71	LINC00351 0.36	ABAT 0.69	FUT4 0.88
FAM117A^{expr} 0.51	RGS20 0.59	CDC42EP3 0.58	FGGY 0.31	E2F7 0.64	TSR1 0.77
CLIC6^{expr} 0.45	FUT4 0.54	FAM160A1 0.52	TM4SF4 0.31	XCR1 0.62	SRR 0.74
FUT4^{expr} 0.41	CLIC6 0.53	HNRNPM 0.49	RN7SL855P 0.31	MLF1IP 0.46	STK17B 0.62
IRX5^{expr} 0.39	IRX5 0.48	NR2F2 0.48	LRP1B 0.29	ERO1L 0.43	PHTF2 0.56
ZFAND2A^{methyl} 0.39	PAOX 0.39	IVD 0.46	ZNF713 0.29	PSMD2 0.43	SEPT2 0.51
PDZD8^{methyl} 0.32	CLEC17A 0.35	MEX3C 0.39	FOCAD 0.28	TFDP1 0.38	C20orf11 0.50
PAOX^{expr} 0.29	TYRP1 0.23	FAM53B 0.33	ZNF733P 0.27	ZNF557 0.32	RFC4 0.48

Table 5. 10 most frequently selected variables of LUAD in different models using bootstrap methods. *expr*: mRNA expression features; *methyl*: methylation features.

1. Fielding LP, Fenoglio-Preiser CM, Freedman LS (1992) The future of prognostic factors in outcome prediction for patients with cancer. Cancer 70: 2367-2377.
2. Michiels S, Koscielny S, Hill C (2005) Prediction of cancer outcome with microarrays: a multiple random validation strategy. Lancet 365: 488-492.
3. Wu TT, Gong H, Clarke EM (2011) A transcriptome analysis by lasso penalized Cox regression for pancreatic cancer survival. J Bioinform Comput Biol 9 Suppl 1: 63-73.
4. Meng L, Xu Y, Xu C, Zhang W (2016) Biomarker discovery to improve prediction of breast cancer survival: using gene expression profiling, meta-analysis, and tissue validation. Onco Targets Ther 9: 6177-6185.
5. Zhao S, Bai N, Cui J, Xiang R, Li N (2016) Prediction of survival of diffuse large B-cell lymphoma patients via the expression of three inflammatory genes. Cancer Med 5: 1950-1961.
6. Luan ZM, Zhang H, Qu XL (2016) Prediction efficiency of PITX2 DNA methylation for prostate cancer survival. Genet Mol Res 15.
7. Han ZB, Chen HY, Fan JW, Wu JY, Peng ZH, et al. (2013) [Expression and survival prediction of microRNA-155 in hepatocellular carcinoma after liver transplantation]. Zhonghua Yi Xue Za Zhi 93: 884-887.
8. Kang MJ, Kim J, Jang JY, Park T, Lee KB, et al. (2014) 22q11-q13 as a hot spot for prediction of disease-free survival in bile duct cancer: integrative analysis of copy number variations. Cancer Genet 207: 57-69.
9. Yu YP, Song C, Tseng G, Ren BG, LaFramboise W, et al. (2012) Genome abnormalities precede prostate cancer and predict clinical relapse. Am J Pathol 180: 2240-2248.
10. Mariani L, Coradini D, Biganzoli E, Boracchi P, Marubini E, et al. (1997) Prognostic factors for metachronous contralateral breast cancer: a comparison of the linear Cox regression model and its artificial neural network extension. Breast Cancer Res Treat 44: 167-178.
11. Zheng QQ, Wang P, Hui R, Yao AM (2009) Prognostic analysis of ovarian cancer patients using the Cox regression model. Ai Zheng 28: 170-172.
12. Ghiandoni G, Rocchi MB, Signoretti P, Belbusti F (1998) Prognostic factors in gastric cancer evaluated by using Cox regression model. Minerva Chir 53: 497-504.
13. Tibshirani R (1997) The lasso method for variable selection in the Cox model. Stat Med 16: 385-395.
14. Kong S, Nan B (2014) Non-Asymptotic Oracle Inequalities for the High-Dimensional Cox Regression via Lasso. Stat Sin 24: 25-42.
15. Ternes N, Rotolo F, Michiels S (2016) Empirical extensions of the lasso penalty to reduce the false discovery rate in high-dimensional Cox regression models. Stat Med 35: 2561-2573.
16. Chen C, Zhang C, Cheng L, Reilly JL, Bishop JR, et al. (2014) Correlation between DNA methylation and gene expression in the brains of patients with bipolar disorder and schizophrenia. Bipolar Disord 16: 790-799.
17. Liu C, Jiang J, Gu J, Yu Z, Wang T, et al. (2016) High-dimensional omics data analysis using a variable screening protocol with prior knowledge integration (SKI). BMC Systems Biology 10: 457.
18. Zou H (2006) The adaptive lasso and its oracle properties. Journal of the American statistical association 101: 1418-1429.
19. Brennan CW, Verhaak RG, McKenna A, Campos B, Noushmehr H, et al. (2013) The somatic genomic landscape of glioblastoma. Cell 155: 462-477.

20. Westcott PM, Halliwill KD, To MD, Rashid M, Rust AG, et al. (2015) The mutational landscapes of genetic and chemical models of Kras-driven lung cancer. Nature 517: 489-492.

21. Tolosi L, Lengauer T (2011) Classification with correlated features: unreliability of feature ranking and solutions. Bioinformatics 27: 1986-1994.

22. Zhao Q, Shi X, Xie Y, Huang J, Shia B, et al. (2015) Combining multidimensional genomic measurements for predicting cancer prognosis: observations from TCGA. Brief Bioinform 16: 291-303.

23. Wang D, Yan L, Hu Q, Sucheston LE, Higgins MJ, et al. (2012) IMA: an R package for high-throughput analysis of Illumina's 450K Infinium methylation data. Bioinformatics 28: 729-730.

24. Fan J, Yang F, Wu Y (2010) High-dimensional variable selection for Cox's proportional hazards model. Statistics 105: 205-217.

25. Zheng Y, Fei Z, Zhang W, Starren JB, Liu L, et al. (2014) PGS: a tool for association study of high-dimensional microRNA expression data with repeated measures. Bioinformatics 30: 2802-2807.

26. Cheney W, Kincaid DR (2013) Linear Algebra: Theory and Applications. Jones & Bartlett Publ.

27. Bender R, Augustin T, Blettner M (2005) Generating survival times to simulate Cox proportional hazards models. Statistics in Medicine 24: 1713-1723.

28. Meng C, Kuster B, Culhane AC, Gholami AM (2014) A multivariate approach to the integration of multi-omics datasets. BMC Bioinformatics 15: 162.

29. Deng M, Brägelmann J, Kryukov I, Saraiva-Agostinho N, Perner S (2017) FirebrowseR: an R client to the Broad Institute's Firehose Pipeline. Database the Journal of Biological Databases & Curation 2017.

30. Li B, Dewey CN (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 12: 323.

31. Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R, et al. (2011) GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. 12: R41.

32. Louhimo R, Hautaniemi S (2011) CNAmet: an R package for integrating copy number, methylation and expression data. Bioinformatics 27: 887-888.

33. Pencina MJ, D'Agostino RB (2004) Overall C as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. Statistics in Medicine 23: 2109-2123.

34. Hofland J, Steenbergen J, Voorsluijs JM, Verbiest MM, de Krijger RR, et al. (2014) Inhibin alpha-subunit (INHA) expression in adrenocortical cancer is linked to genetic and epigenetic INHA promoter variation. PLoS One 9: e104944.

35. Laurila E, Savinainen K, Kuuselo R, Karhu R, Kallioniemi A (2009) Characterization of the 7q21-q22 amplicon identifies ARPC1A, a subunit of the Arp2/3 complex, as a regulator of cell migration and invasion in pancreatic cancer. Genes Chromosomes Cancer 48: 330-339.

36. Liu Z, Yang X, Chen C, Liu B, Ren B, et al. (2013) Expression of the Arp2/3 complex in human gliomas and its role in the migration and invasion of glioma cells. Oncol Rep 30: 2127-2136.

37. Shen R, Olshen AB, Ladanyi M (2009) Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. Bioinformatics 25: 2906-2912.

38. Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, et al. (2014) Similarity network fusion for aggregating data types on a genomic scale. Nature Methods 11: 333.

39. Vaske CJ, Benz SC, Sanborn JZ, Earl D, Szeto C, et al. (2010) Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. Bioinformatics 26: i237-i245.

40. Bergersen LC, Glad IK, Lyng H (2011) Weighted lasso with data integration. Stat Appl Genet Mol Biol 10: 1-29.

41. Tipping ME, Bishop CM (1999) Probabilistic Principal Component Analysis. Journal of the Royal Statistical Society 61: 611–622.

42. Bastien P (2004) PLS-Cox model: application to gene expression. Compstat —proceedings in Computational Statistics: 655-662.